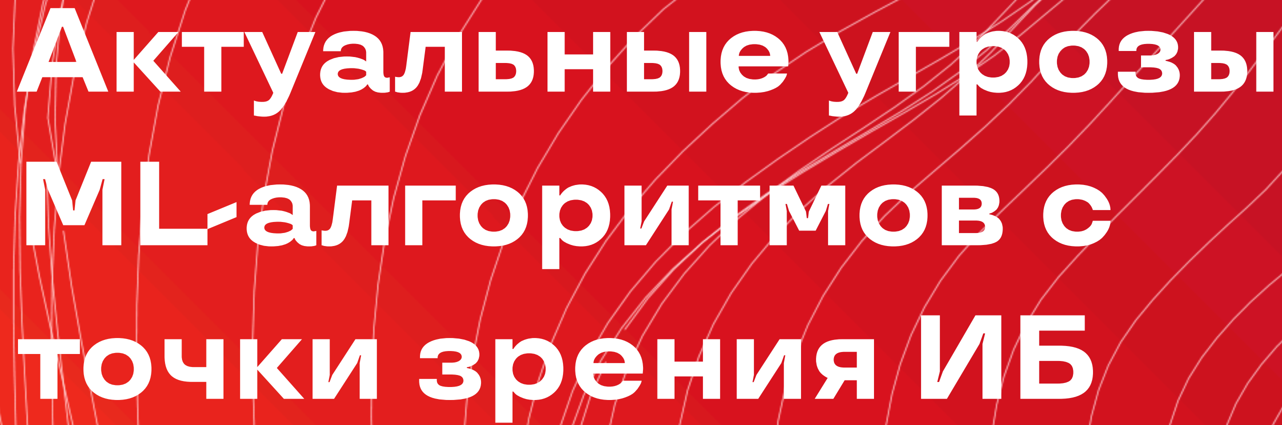


Актуальные угрозы ML-алгоритмов с точки зрения ИБ

Александра Мурзина, Positive Technologies



ML применяется в ИБ

- Для детектирования атак
  - Продукты
  - Экспертиза
- Как инструмент хакера
- Сам может быть уязвим



Есть ли проблема?

- Шифровальщики
- Сливы данных
- Ну дипфейки, но выглядит как шутка какая-то

Есть ли что-то, что приносит ущерб?



ML в нашей жизни

- Биометрия
- Рекомендательные системы
- Переводчики
- Системы анализа спама и фрода
- Голосовые ассистенты
- IoT, умные устройства



ML в нашей жизни

Можно ли представить жизнь без этих сервисов?

А как мы знаем, вопросы ИБ актуальны как никогда.



1. Какие атаки изучают?

2. Чего боятся вендоры?

3. Что происходит на самом деле?



Какие атаки изучают больше всего

Бывает, что все атаки на МЛ, которые есть, обобщают до Adversarial attack.



Какие атаки изучают больше всего

Бывает, что все атаки на МЛ, которые есть, обобщают до Adversarial attack.

Всегда найдутся данные, на которых алгоритм ошибется.
Задача злоумышленника – подобрать такие данные.
Как?



Какие атаки изучают больше всего

Бывает, что все атаки на МЛ, которые есть, обобщают до Adversarial attack.

Всегда найдутся данные, на которых алгоритм ошибется.
Задача злоумышленника – подобрать такие данные.
Как?

С помощью adversarial техник найти тот самый blind spot модели – такое место, где модель не уверена в принимаемом решении.



Adversarial attacks in the wild

Затруднение копирования контента

- Авито публиковали контент с замененным номерным знаком
- Авто.ру копировали, заменяя номерной знак на свой



https://tinyurl.com/36348wd4



Adversarial attacks in the wild

Затруднение копирования контента

- Авито публиковали контент с замененным номерным знаком
- Авто.ру копировали, заменяя номерной знак на свой

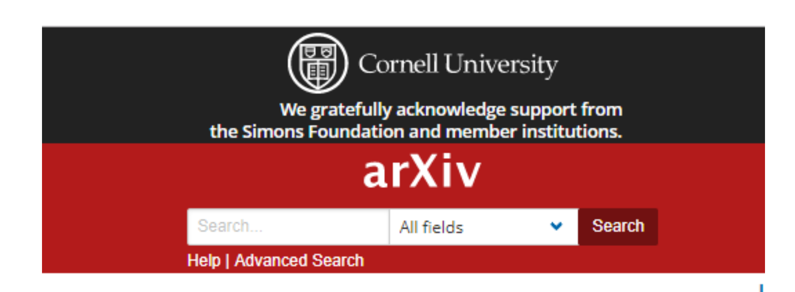
- Авито готовили изображение, использовав Adversarial attack
- ML-модель Авто.ру больше не могла обнаруживать номерной знак и менять его на свой



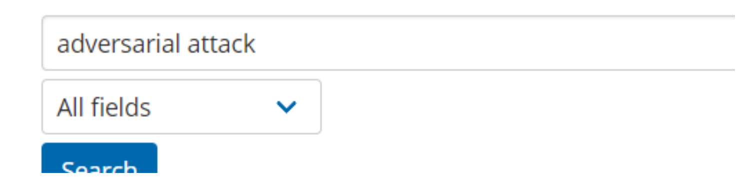
https://tinyurl.com/36348wd4

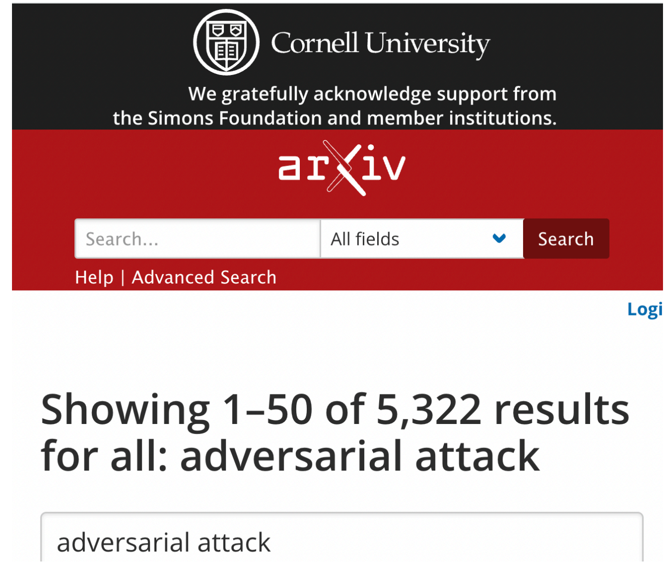


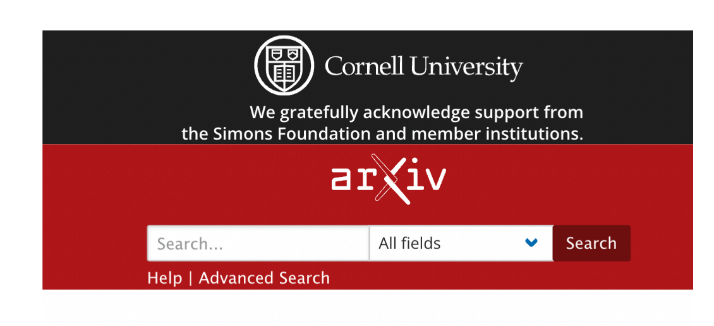
Какие атаки изучают больше всего



Showing 1–50 of 1,504 results for all: adversarial attack







Showing 1–50 of 6,161 results for all: adversarial attack

adversarial attack

All fields

1,504	5,322	6,161
2019	**апрель 2022**	**ноябрь 2022**



Риски безопасности ИИ от Kang Li

AI App Security Risk

Model Security

- **Adversarial ML**
- Model Backdoor
- Model Theft

Implementation Security

- Sensor Security
- Flaws in Framework
- Logical Flaws

Data Integrity Security

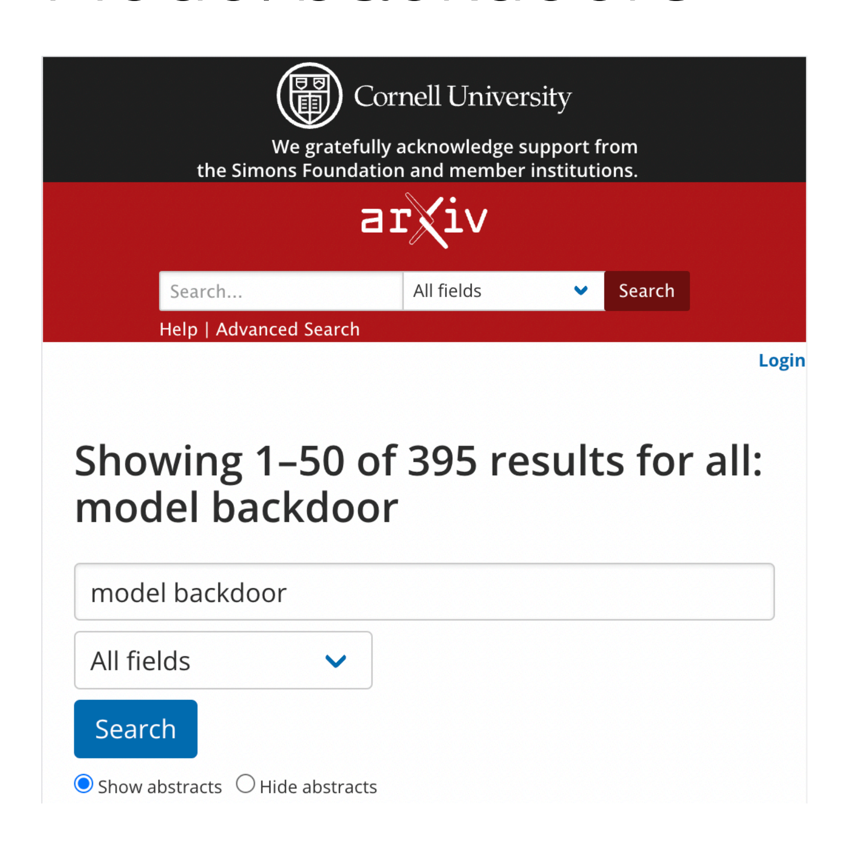
- Data Poisoning
- Scaling Attack
- Risk over Network

https://tinyurl.com/4fh7j3ky



Какие атаки изучают больше всего

Model backdoors

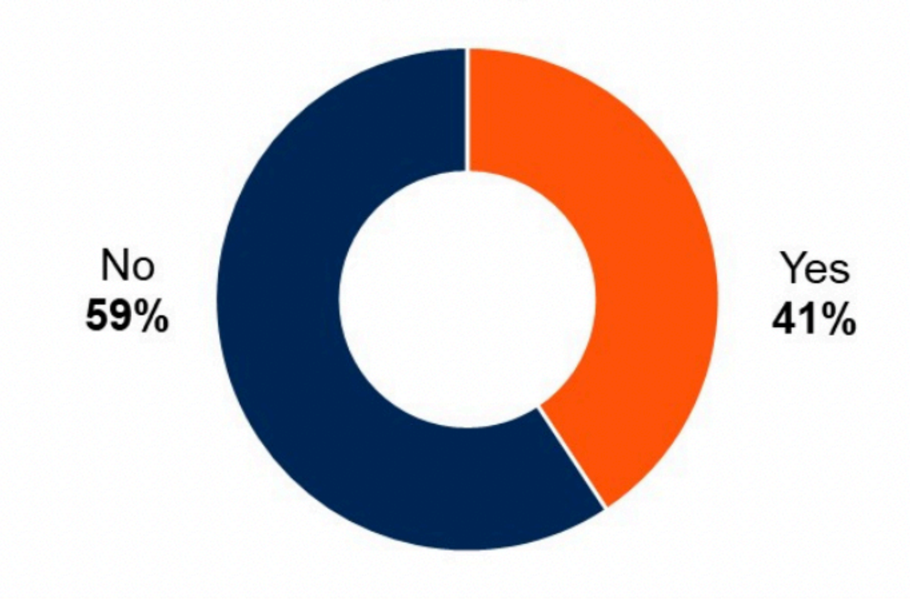


— так какая атака опасней?
— та, которая причиняет наибольший ущерб
— так какая?
— а какие у вас риски?



Чего боятся вендоры? Обзор Gartner

Случались ли у вас нарушения конфиденциальности ИИ или инциденты безопасности?



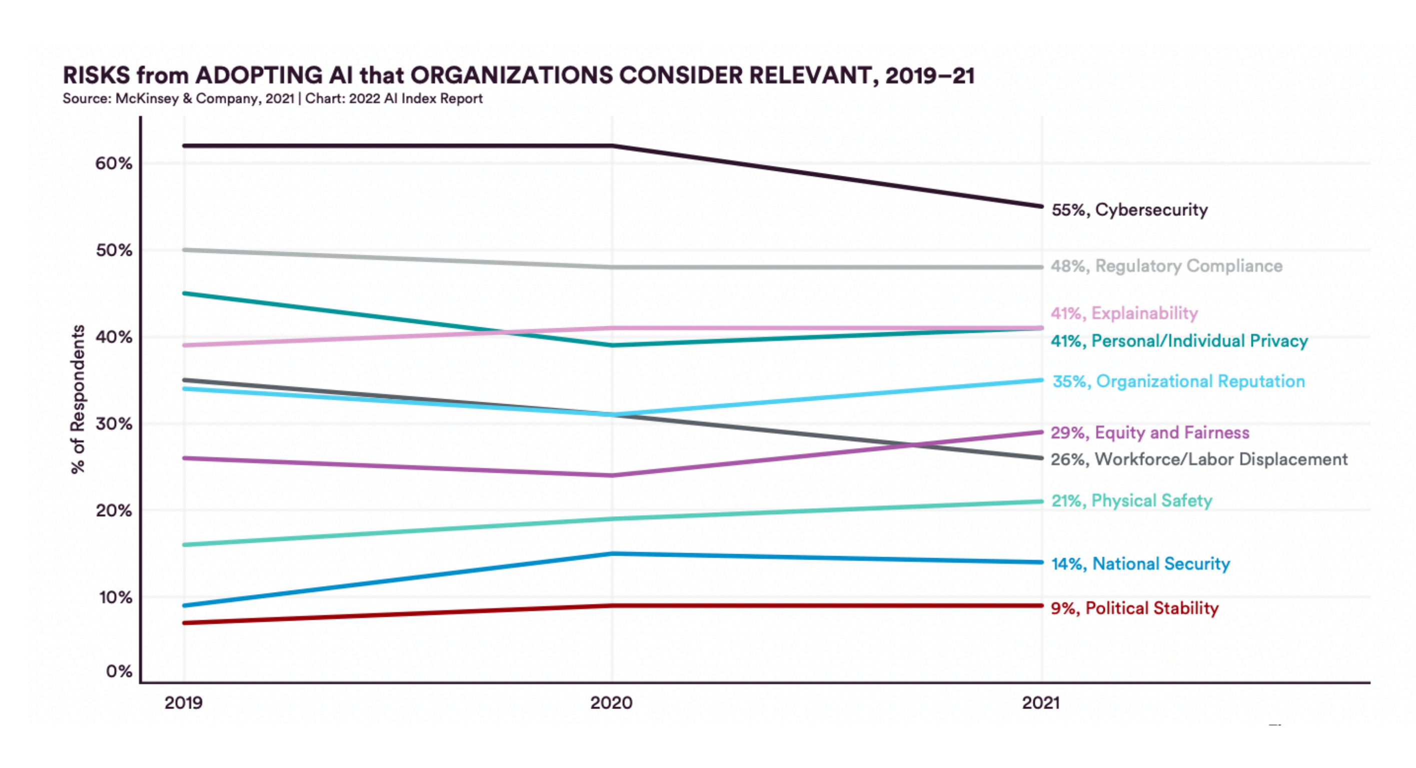
https://tinyurl.com/mr3fhkvx

41% опрошенных организаций столкнулись с с нарушением конфиденциальности ИИ или инцидентом безопасности.

Из этих инцидентов 60% были компрометацией данных внутренней стороной, а 27% — злонамеренными атаками на ИИ-инфраструктуру организации.



Есть ли проблема?



Наиболее актуальным риском в 2021 году является кибербезопасность (55% респондентов).

Далее следуют нормативно-правовое соответствие (48%), интерпретируемость (41%) и конфиденциальность (41%).

Consideration and Mitigation of Risks From Adopting AI

https://tinyurl.com/3kax9hbw



Что происходит на самом деле?

спойлер:

нет официальной статистики по инцидентам



Сделали копию лица для DeepFake-видео

Неизвестные скопировали лицо основателя Dbrain
Дмитрия Мацкевича для deepfake-видео с рекламой супердоходов.



https://tinyurl.com/2x7755vw



Оформить кредит в банке по биометрии?

29 марта 2021

«Компания расширит функции запущенного в 2018 году приложения «Биометрия» услугой дистанционной сдачи биометрии. Это предложение рассматривается ФСБ и Федеральной службой по техническому и экспортному контролю (ФСТЭК)»

https://tinyurl.com/yepaz9r9



Оформить кредит в банке по биометрии?

29 марта 2021

«Компания расширит функции запущенного в 2018 году приложения «Биометрия» услугой дистанционной сдачи биометрии. Это предложение рассматривается ФСБ и Федеральной службой по техническому и экспортному контролю (ФСТЭК)»

https://tinyurl.com/yepaz9r9

10 апреля 2021

В Москве мошенники стали использовать голоса клиентов банков для оформления кредитов.

https://tinyurl.com/ypyhje2y



Можно ли манипулировать ML-системой?

Полиция в США включает популярную музыку на забастовках, чтобы исключить попадание видео с собой в соц сети.

> “You can record all you want, I just know it can't be posted to YouTube.

https://tinyurl.com/5b9suxjy



А нас это как касается?





IoT & биометрия

Исследования Positive Technologies:

- Умный дверной звонок
- Устройство биометрической идентификации



Умный дверной звонок

Google
cameras with facial recognition AI
Line crossing, Intrusion in area, Loitering, People gathering. DSE buy ...

www.nature.com › news feature
Resisting the rise of facial recognition - Nature
18 Nov 2020 — Facial-recognition technology (FRT) has long been in use at airport borders ...
Serbia an inappropriate place for implementation of AI cameras.

People also search for

Surveillance Cameras For Facial Recognitions
Tend Secure Lynx Indo...
Nest Cam IQ Outdoor
Nest Cam IQ Indoor
Netatmo Welcome
See more

A Clear View
Une vision claire
Ein klares Bild
Vista chiara
Vista nítida
Een helder beeld
Uma Visibilidade Nítida

Human Detection
Détection humaine
Personenerkennung
Rilevamento della presenza di esseri umani nei paraggi
Detección de personas
Menselijke detectie
Deteção Humana

See More
En savoir plus
Mehr sehen
Visualizza maggiori dettagli
Visualización ampliada
Zie meer
Veja Mais

Facial snapshot notifications
Notifications par instantané facial
Benachrichtigungen mit Gesichtsaufnahmen
Notifiche con istantanea dei volti
Notificaciones de instantáneas faciales
Meldingen met snapshot van gezicht
Notificações de instantâneos faciais

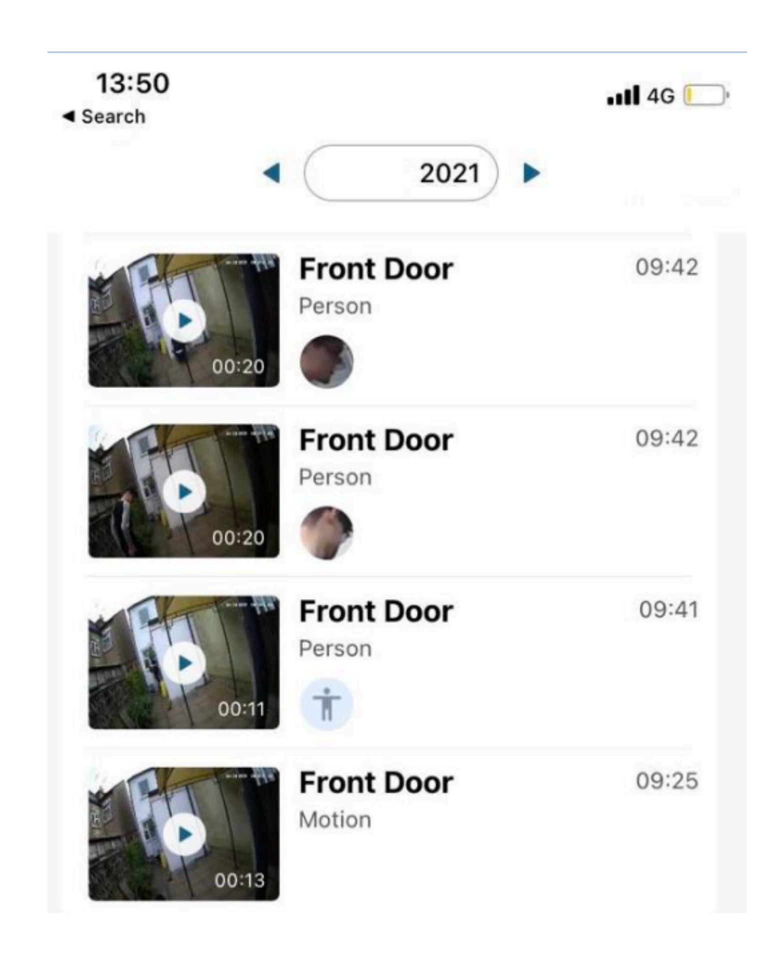
IP65 Weatherproof
Résistant aux intempéries IP 65
Wettergeschützt nach IP65
Impermeabilità IP65
Clasificación de impermeabilidad IP65
IP65-weerbestendig
Impermeabilidade IP65

"Hey Google, show me my front door"
« Eh, Google, montre-moi le seuil de la porte »
„Hey Google, zeig mir die Haustür"
"Hey Google, mostrami l'entrata"
"Ok Google, muéstrame la puerta de entrada"
Hey Google, toon mij mijn voordeur
"Hey Google, mostra-me a porta da frente"

- Вендор заявляет о безопасности.
- Может открывать замок автоматически.
- Может быть подключен к общему хабу умного дома.



Умный дверной звонок



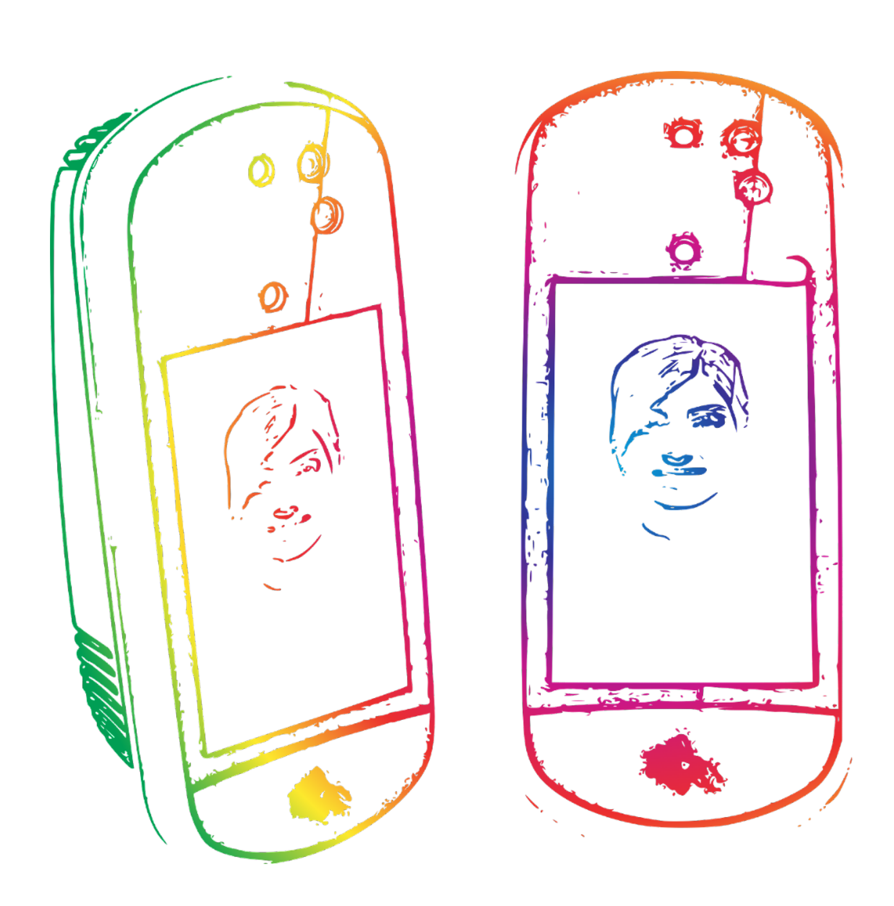
- Вендор заявляет о безопасности.
- Может открывать замок автоматически.
- Может быть подключен к общему хабу умного дома.

- Слабое шифрование на этапе авторизации.
- Возможность получить доступ к хабу умного дома.
- Можно следить за тем, кто приходит/уходит.

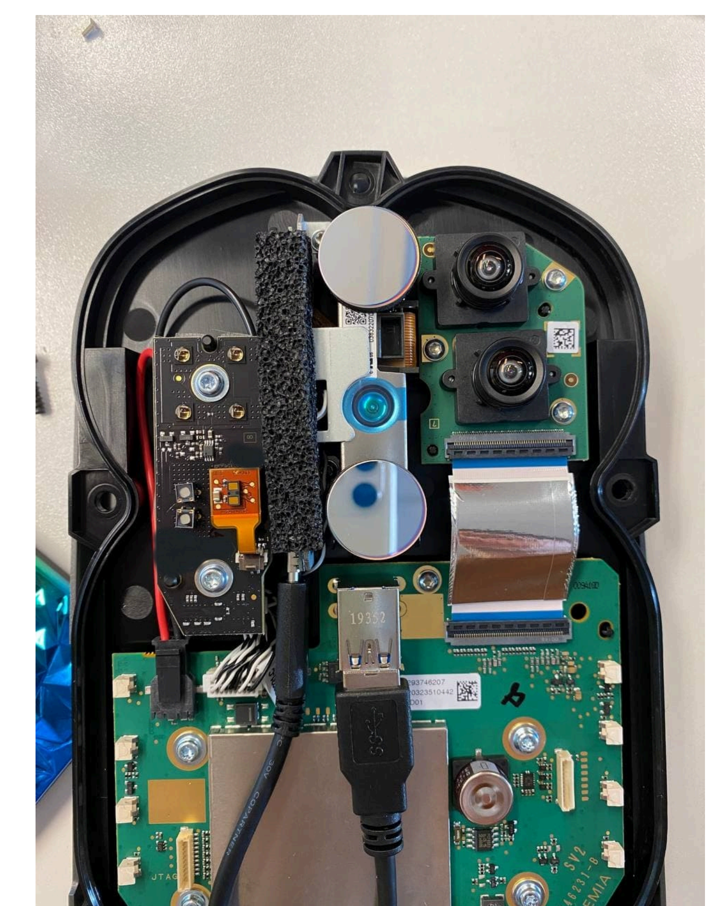




Умное устройство биометрической идентификации



- Камера глубины
- 2 камеры видимого диапазона



https://youtu.be/2fvdyKl1NiY





Умное устройство биометрической идентификации

Как работает?

1. Детектируем лицо в кадре.
2. Проверяем Liveness камерой глубины.
3. Захватываем лицо из камеры видимого диапазона.
4. Предобработка.
5. DNN.
6. Сравнение с БД.

https://youtu.be/2fvdyKl1NiY

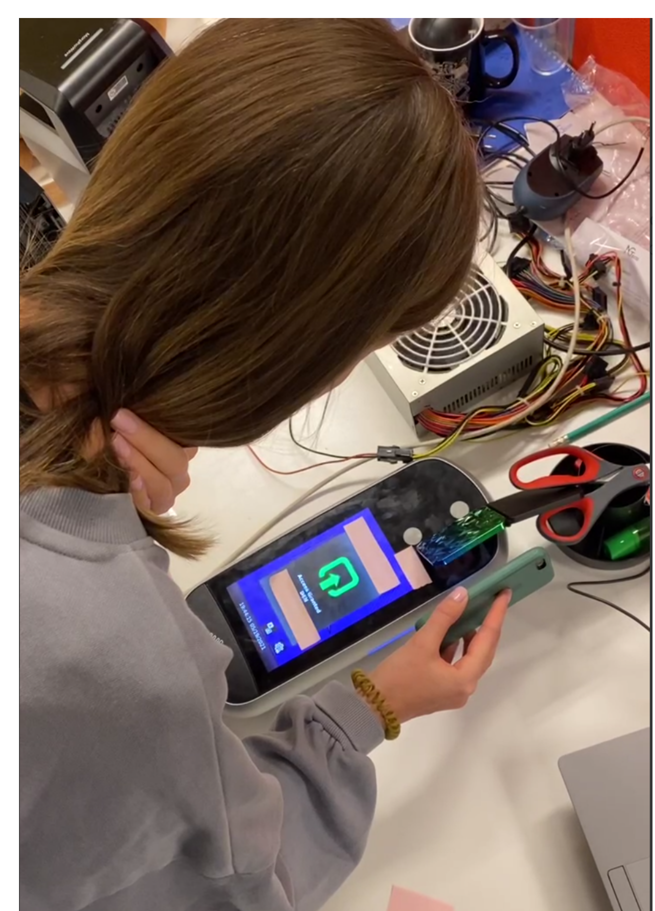


Умное устройство биометрической идентификации

Как работает?

1. Детектируем лицо в кадре.
2. **Проверяем Liveness камерой глубины.**
3. **Захватываем лицо из камеры видимого диапазона.**
4. Предобработка.
5. DNN.
6. Сравнение с БД.

https://youtu.be/2fvdyKl1NiY



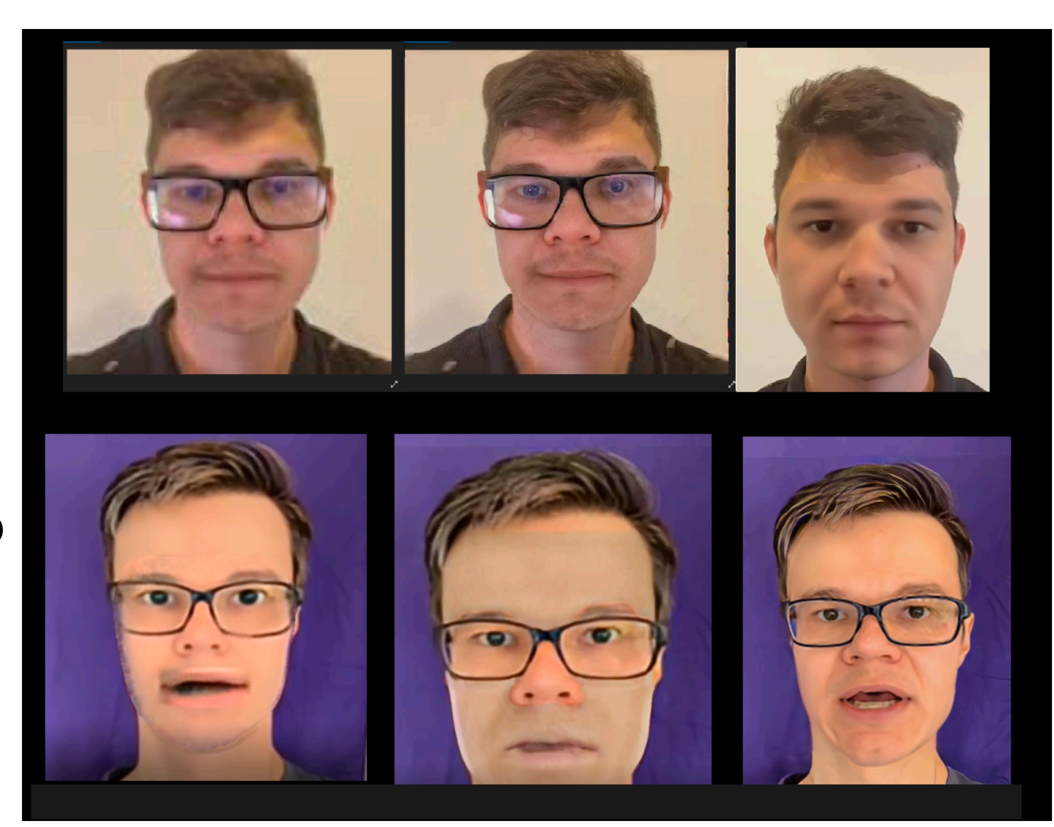




KYC

KYC (Know Your Customer) — обязательная для финансовых институтов процедура идентификации контрагентов. Включает в себя процедуры отбора и идентификации, а также отслеживание транзакций и их анализ.

- Онлайн — наше всё.
- Но надо как-то подтверждать свою личность и далее транзакции.



https://tinyurl.com/2s4vhkvj



Что делать?

Зарегулировать!

Совфед запретит искусственному интеллекту дискриминировать россиян

The EU wants to put companies on the hook for harmful AI

https://tinyurl.com/ykc56tkr
https://tinyurl.com/3h7526hz



Может, не надо?

George Hotz, основатель стартапа Comma.AI:

> Today's struggle isn't a technological one, it's psychosocial. The technology of sovereignty is only part of the battle, particularly if that technology is further down the tech tree than non invasive wireheading.

Comma.AI — опенсорс-ассистент водителя с поддержкой 200+ машин — опенсорс-беспилотник.

https://tinyurl.com/ccczpfd8



Модель угроз

Raw data

API

Cloud (training)

dataset assembly

- engineering
- training
- validation
- testing

model

API

Backend (deployment)

ML service

model

output

1 2 3 4 5 6 0

0 — api security

1 — poisoning

2 — trojaning

3 — backdooring

4 — model extraction

5 — model inversion

6 — inference attacks



И что с этим делать?

- Багбаунти
- Киберполигоны
- Внешний аудит

https://tinyurl.com/mut5d6z7
https://tinyurl.com/42vcnpzp



Lessons learned

- То, что изучают, - не всегда то, чего боятся вендоры и не всегда то что в происходит в реальной жизни.
- Важно понимание угроз, которые кажутся наиболее опасными.
- Стоит начать с классической безопасности.
- Некоторых рисков ИБ в МЛ можно избежать, если учесть в архитектуре решения.
- Это бесконечная игра.



за доклад можно голосовать тут



Актуальные угрозы ML-алгоритмов с точки зрения ИБ

Александра Мурзина, Positive Technologies

@murzina_a

